# Factors influencing estimates of coordinate error for molecular replacement

Authors
**Kaushik S. Hatti[a], Airlie J. McCoy[a], Robert D. Oeffner[a], Massimo D. Sammito[a] and Randy J. Read[a]***
[a]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK
Correspondence email: rjr27@cam.ac.uk

**Synopsis** We propose improved coordinate error estimates for X-ray and NMR models used for maximum-likelihood based molecular replacement phasing.

**Abstract** Good prior estimates of the effective root-mean-square deviation (RMSD) between atomic coordinates of the model and the target optimise the signal in molecular replacement, thereby increasing the success rate in difficult cases. Previous studies using protein structures solved by X-ray crystallography as models showed that optimal error estimates (refined after structure solution) were correlated with sequence identity between the model and target, and with the number of residues in the model. Here we have extended this work to find additional correlations between parameters of the model and target and hence improved prior estimates of the coordinate error. Using a graph database, a curated set of 6,030 molecular replacement calculations using models that had been solved by X-ray crystallography was analysed to consider about 120 model and target parameters. Improved estimates were achieved by replacing the sequence identity with the Gonnet score for sequence similarity, as well as by considering the resolution of the target structure and the *MolProbity* score of the model. This approach was extended by analysing 12,610 additional molecular replacement calculations where the model was determined by NMR. The median RMSD between pairs of models in an ensemble was found to be correlated with the estimated RMSD to the target. For models solved by NMR, the overall coordinate error estimates were larger than for structures determined by X-ray crystallography, and were more highly correlated with the number of residues.

**Keywords: Molecular replacement; coordinate error; RMSD; NMR; LLG**

## 1. Introduction

Likelihood-based molecular replacement (MR) uses estimates of the errors in the model and data to improve the signal to noise in the search. In Phaser (McCoy *et al.*, 2007), the Log Likelihood Gain on Intensities (LLGI) (Read & McCoy, 2016) accounts for the effect of intensity measurement errors when scoring MR searches. The LLGI discriminates correct from incorrect solutions and is used to rank solutions across complex search strategies (Oeffner *et al.*, 2018), such as those implemented in the

ARCIMBOLDO suite of programs (Millán *et al.*, 2015), AMPLE (Rigden *et al.*, 2008; Bibby *et al.*, 2013) and MrBUMP (Keegan & Winn, 2008).

The LLGI (for acentric reflections) is defined in equation (1):

$$LLGI = \sum_{hkl} log\left[\frac{2E_e}{1 - D_{obs}^2\,\sigma_A^2}\; exp\left(\frac{E_e^2 + D_{obs}^2\,\sigma_A^2 E_c^2}{1 - D_{obs}^2\,\sigma_A^2}\right) I_o\left(\frac{2E_e D_{obs}\,\sigma_A E_c}{1 - D_{obs}^2\,\sigma_A^2}\right)\right] \qquad (1a)$$

$$\sigma_A = \sqrt{f_p}\; exp\left(\frac{-2(\pi\, s\, \Delta)^2}{3}\right) \qquad (1b)$$

In this equation, the parameters $E_e$ (effective E) and $D_{obs}$ (Luzzati-style D factor) are derived from the measured intensity and its estimated standard deviation (Read & McCoy, 2016), resulting in any reflections with large experimental errors being downweighted. This gives an excellent approximation to an intensity-based likelihood target that would require expensive numerical integration. The $\sigma_A$ term accounts for the effect of predicted errors in the model. LLGI calculations will be optimal when the initial estimates of $\sigma_A$ are accurate. Underestimation of $\sigma_A$ will lead to underweighting of the high-resolution reflections in the LLGI calculations, whereas overestimation of $\sigma_A$ will lead to overweighting of these reflections. Both problems will lead to sub-optimal usage of data and can influence success in a borderline case.

Ignoring an optional bulk-solvent term for simplicity, $\sigma_A$ can be expressed as a function of resolution ($s=1/d$), model completeness ($f_p$, the fraction of total scattering accounted for by the model) and the effective RMS coordinate error of the model ($\Delta$) as given in equation (1b). Once the model has been placed in the MR calculation, the value of $\Delta$ can be refined during a rigid-body refinement. This term $\Delta$ is different from the RMSD that can be calculated between equivalent atomic positions by superposing two structures, because it is an effective RMSD that optimises the variance term in the LLGI target. For this reason, we refer to it as variance-RMSD or in short VRMS.

The VRMS can only be refined once a model has been placed, and its value is only relevant if the model is placed correctly, so it is necessary to provide a prior estimate of VRMS before carrying out the search. Prior to *Phaser* version 2.5.4, *Phaser* used the Chothia & Lesk (1986) curve (which relates sequence identity to the RMSD between main chain atoms) as a first-order approximation. Although these values worked reasonably well, it became clear that estimates tailored to the MR problem were needed. We developed an improved functional form to estimate VRMS (equation (2) as a function of size of the model (*Nres*) and sequence identity (*H*: fraction of mutated residues) between model and the target (Oeffner *et al.*, 2013).

$$eVRMS = A\,(B + Nres)^{\,1/3}\; \exp(CH) \qquad (2)$$

However, experience using a wide variety of MR models showed that sequence identity is a poor measure to assess sequence similarity of very distant homologues. We considered a number of alternative sequence similarity measures that have been developed over the past few decades, summarised very well by Vogt *et al.,* (1995).

To assess which property might improve predictive power, we also investigated correlations of a variety of properties of the model and the target with the refined VRMS term. Because the work up to this point had concentrated on models derived by X-ray crystallography, we also developed a new functional form to estimate VRMS specifically for members of NMR ensembles used as phasing models.

## 2. Methodology

The study follows the methods described by Oeffner *et al* ( 2013). Here, we summarise the steps from Oeffner *et al* used in carrying out large scale molecular replacement trials for X-ray models. The extension of the earlier work to include NMR models is elaborated below.

### 2.1. Generation of molecular replacement data using X-ray models

In the earlier study, a total of 2862 structures (and associated diffraction data) with a single chain in the asymmetric unit, across a range of SCOP classes (Murzin *et al.*, 1995) and with size varying between 50 and 1500 residues, were selected as targets from the wwPDB (Berman *et al.*, 2000). Care was taken not to include targets that were known to be twinned or for which the published R factors could not be reproduced by the Uppsala Electron-Density server (Kleywegt *et al.*, 2004). Only one example was kept for each unique sequence, except that all entries for proteins with more than 600 residues were retained to improve sampling of large targets. For each target, homologous structures were identified by performing a BLAST search (Altschul, 1991) with the BlastP tool against the wwPDB. ClustalW (Thompson *et al.*, 1994) was used to perform pair-wise alignments of the homolog and target sequences; unlike BLAST, which finds local subsequence alignments, ClustalW maximises the global sequence alignment. The models were pruned and edited with Sculptor (Bunkóczi & Read, 2011*a*). A total of 21,822 molecular replacement calculations was performed and used for the analysis in the earlier study.

For this study, we curated the database from the earlier study to remove redundant targets (inadvertently included more than once) and models that failed to lead to successful molecular replacement solutions. To measure the reliability of the molecular replacement solution, we calculated model to map correlations (globalCC) using phenix.get_cc_mtz_pdb to assess the agreement between $2mF_O$-$DF_C$ maps (Read, 1986) computed from the molecular replacement solution and the deposited model. A subset of 6030 molecular replacement trials with globalCC > 0.2 was chosen, in the end, for the curated database. These trials arise from a combination of 1307 distinct targets (which includes 119 targets with deposited intensity data) and 3420 distinct models. The database was extended to include a variety of parameters associated with target, model and sequence similarity measures.

**Target properties**: Several measures to assess crystal parameters, data parameters and protein parameters were downloaded from the wwPDB. See Table 1 for a complete list of target properties considered in the study.

**Model properties**: Parameters such as number of residues, date of deposition, resolution, RMS deviations of bond lengths and angles from ideal values and R-factors were downloaded from the wwPDB. Validation parameters such as Ramachandran properties, clashscore, rotamer outliers, *MolProbity* score (Chen *et al.*, 2010) and Cβ-deviations were recalculated for the processed models using PHENIX (Adams *et al.*, 2010) command-line tools. Non-sphericity of the model was estimated by calculating principal axes using Gromacs (Abraham *et al.*, 2015) command-line tools.

When available, SCOP definitions were downloaded from the SCOPe database (Fox *et al.*, 2014) and assigned to both target and model entries (Table 1).

**Sequence similarity properties:** Several amino acid substitution matrices were used to assess sequence similarity of a target-model pair. In this study, we considered matrices that were judged to assess sequence similarity accurately for pair-wise sequence identities below 50% (Vogt *et al*., 1995) (Table 1). The matrices were used from within Biopython (version 1.72) to score every target-model pairwise sequence alignment. The scores were normalised for the length of aligned residues.

### 2.2. Generation of molecular replacement data using NMR models

A protocol similar to that used to generate molecular replacement data with X-ray models was used for NMR models. Targets identified above were retained and no new targets were considered for this study.

**Selection of NMR models**: The sequence profile database constructed using the entries from the PDB at 70% sequence non-redundancy, PDB_mmCIF70, was downloaded from the HHpred (Zimmermann *et al.*, 2018) website. For a given target (as selected previously in section 2.1) HMMER (Finn *et al.*, 2011) was used to identify homologous structures from the PDB_mmCIF70. 1364 homologous structures which were determined through NMR alone were retained. Properties specific to NMR models such as number of models deposited in an ensemble and chemical shift data validation were downloaded from the wwPDB data (if reported).

**Processing of NMR models:** Clustal-Omega (Sievers *et al.*, 2011), an improved implementation of the Clustal algorithm, was used to perform pairwise alignment of target and NMR model sequences. The scores discussed for X-ray models were also used to evaluate sequence similarity for NMR models. Models were pruned and edited with Sculptor (Bunkóczi & Read, 2011*a*). Other studies have shown that using NMR models for MR phasing is a challenge and suggested trimming protocols to improve success in molecular replacement phasing (Chen *et al.*, 2000; Mao *et al.*, 2011). Accordingly, ensembles were generated with Ensembler (Bunkóczi & Read, 2011*b*), selecting the default option to trim residues deviating by more than 3Å. Gesamt (Krissinel, 2012) was used to perform a pair-wise combination

superposition of all – vs – all trimmed models in an NMR ensemble. A median RMSD between equivalent C$\alpha$ positions was calculated for each trimmed ensemble to assess the conformational differences among the models. See Table 1 for the list of NMR specific metrics considered in this study.

**Molecular replacement rigid-body refinement:** NMR models with over 50% coverage were superposed onto the target using Gesamt. A total of 20,973 molecular replacement rigid-body refinements was performed using the MR_RNP mode of *Phaser* (McCoy *et al.*, 2007) using each model from the trimmed NMR ensemble independently. In practice, it is best to use NMR models as ensembles, but success in statistical weighting of the ensembles depends on having the best estimate of the effective error of each individual member of the ensemble (Read, 2001).

### 2.3. Generation of Graph database

For a given pair of target and model, there were about 120 properties to be evaluated. To address this large-scale comparison, we built an in-house database representing the data as a graph, using the open-source graph database platform Neo4j (Version 3.4.0; URL: https://neo4j.com). Target and model were defined as nodes and an edge connecting the two defined a relationship (Figure 1A). All the properties associated with a target or a model were associated with their respective nodes. Properties such as sequence similarity scores and results of molecular replacement calculations were associated with the edge connecting the two nodes. In this way, a complex graph network was generated, which included all the data defining the targets, models (both X-ray and NMR) and the relationships between them (Figure 1B). An intermediary layer of nodes (not shown in Figure 1 for the sake of clarity) was used to represent model number in the case of NMR ensembles. Cypher, a declarative graph querying language, was used to query the data.

All statistical analysis was performed within the R statistical programming environment (R version 3.5.0) (R Core Team, 2018). Non-linear least squares fitting was performed using the *nls* package (Baty *et al.*, 2015) starting with the most highly correlated parameter and subsequently adding more parameters until a low residual correlation with unused parameters was obtained. Figures were generated using the *ggplot2* package (Wickham, 2016), both available within R.

### 2.4. Derivation of equations to predict refined VRMS

In fitting the two data sets, the data were examined to determine which properties were most highly correlated with the refined VRMS. In general, one property was included at a time. Different functional forms were tested for equations adding that property when fitting to the data, and the functional form that minimised the deviation between refined and estimated VRMS was chosen. To choose the next property to include in the fit to the data, residual correlations (correlation to the normalised difference between the refined and estimated VRMS) were computed. The process was terminated when adding a new property had little effect on the quality of the fit.

## 3. Results

### 3.1. Improved estimates for X-ray models

The Gonnet matrix score (Gonnet *et al*., 1992) has the highest correlation to the refined VRMS term (Table 2) among all the metrics used to estimate sequence similarity, so this was chosen to play the role taken by sequence identity in equation (2) from Oeffner *et al.* ( 2013). Among the properties of the model, the size of the model has the highest correlation to VRMS, followed by the *MolProbity* score. As judged by the residual correlation (also shown in Table 2), the *MolProbity* score was the most significant model feature that had not been considered in the work by Oeffner *et al.* ( 2013). Although we had only expected properties involving the model to play a significant role, we found target resolution also to correlate with VRMS, with a higher correlation than *MolProbity* score (Table 2). Further molecular replacement calculations were performed to ascertain that the correlation is not an artefact of the resolution of data used during VRMS refinement. Molecular replacement calculations were repeated as a function of target resolution by truncating the data to lower resolution limits (2.2Å, 2.7Å, 3.0Å, 3.5Å, 4Å, 6Å, 7Å), only to find that the correlation between VRMS and the original resolution of the target persisted.

Different functional forms for a non-linear least-squares fit to the data from the 6030 molecular replacement trials in the curated database were tested in preliminary work, including sums and products involving different properties and different choices of exponent for terms related to particular properties. The best results were obtained by equations expressing the total variance as a sum of independent variance terms.

Figure 2 shows the effect of including successive variance terms. Diminishing returns were achieved as new properties with lower explanatory power were added. After the *MolProbity* score had been included, the most significant remaining property was the percentage of beta-sheet in the model, with a residual correlation of -0.13. However, including this property in the non-linear fit had very little effect on the quality of fit, so it was not included in the final equation (3). Note that much of the correlation with alpha-helix content had apparently been accounted for by this point by correlations with other properties.

$$eVRMS = \sqrt{A\,(Nres) + B\exp\,(CG^{2.5}) + D\,(MolProbity) + E\,(Resolution)^3} \qquad (3)$$

The non-linear least squares fit of equation (3) yielded the coefficients A=0.001455, B=1.710, C=-0.2444, D=0.1040, E=0.01586. Residual correlations computed using the new expression for eVRMS show that this functional form accounts for most of the initial systematic variation in the data (Table 2). In addition, a frequency distribution computed from the ratios of estimated and refined VRMS values

became more symmetrical and unimodal than using the previous *Oeffner* coordinate error estimate (Figure 3).

Figure 3 also shows that the VRMS distributions are slightly different for different SCOP fold classes, with errors being slightly underestimated on average for all alpha proteins and slightly overestimated for all beta proteins. However, in keeping with the very minor effect on the fit of including percentage beta-sheet content, the differences in the distributions for fold classes are small compared to the width of the overall distribution.

### 3.2. Estimates for NMR models

Previous published work (Chen *et al.*, 2000) and anecdotal evidence suggested that models obtained using NMR data generally work more poorly for MR than models obtained using X-ray data. In addition, we anticipated that a different functional form might be needed to predict model quality. For instance, considering that NMR structures are defined primarily by short-range distance data, one might expect an increased dependence of coordinate error on model size. In addition, NMR structures are usually reported as an ensemble of alternative models (typically 20) that all have comparable fit to the data, and one might expect the deviation among these models to provide an indication of model precision, if not accuracy. Indeed, the analysis of correlations revealed that, for NMR models, there was a stronger correlation between refined VRMS and model size than for X-ray data, and there was a significant correlation with the deviation among the models in the ensemble (Table 3).

We wanted to check if the estimates for NMR models could be improved by including criteria recommended by the NMR validation task force (Montelione *et al.*, 2013). For example, completeness refers to the percentage of chemical shifts that have been assigned. Surprisingly no correlation was found between this completeness measure and VRMS. Other measures were reported only for a fraction of the NMR models included in this study and hence could not be studied further. It may be worth revisiting this analysis when larger numbers of NMR structures report these validation metrics.

A new functional form, given in equation (*4*), was defined, again estimating the overall variance as a sum of independent variance contributions and testing different exponents for the underlying variables. The quality of fit was only weakly affected by the exponent for the *Nres* term, probably because the range of model sizes is limited for NMR models. Unexpectedly, an exponent of 1/3 was slightly better than the exponent of one found for the X-ray fit; even though VRMS is more sensitive to model size for NMR compared to X-ray models, this sensitivity comes from the multiplicative factor *A* rather than the exponent.

$$eVRMS = \sqrt{A\,(Nres)^{(1/3)} + B\;exp(CG) + D\,(MolProbity) + E(Resolution) + F(Median\_RMSD)} \qquad (4)$$

The six parameters in this equation were fit using a subset of 12,610 molecular replacement cases (with globalCC>0.2) where NMR structures were used as models, limiting the data to structures that were

between 30 and 300 residues in length. The *MolProbity* score for equation (*4*) corresponds to the individual *MolProbity* score of each model in a given NMR ensemble. Median RMSD is the median of RMSDs of all pair-wise superpositions of members of a given NMR ensemble. The non-linear least-squares fit yielded the coefficients A=0.4240, B=-1.259, C=0.07804, D=0.1442, E=0.2364, F=0.4130. All residual correlations were close to zero, giving a substantial improvement over *Oeffner* estimates derived from X-ray models (Table 3).

### 3.3. The importance of accurate VRMS estimates

It is important to start the calculations with accurate estimates of VRMS to achieve the highest initial LLGI scores, because the absolute value of the LLGI score is highly correlated to the signal-to-noise achieved in the search (McCoy *et al.*, 2017). To evaluate this, we calculated the LLGI in rigid-body refinements starting with the correctly placed model but without refining the VRMS parameter. The same set of cases used for curve fitting of both X-ray and NMR models were considered in this study. The calculations using both X-ray-derived and NMR-derived models were performed with both the *Oeffner* and new estimates of VRMS. For NMR models, only the first member of the NMR ensemble was considered in these calculations.

An incremental improvement was observed in the case of X-ray models. The LLGI calculated with the new VRMS estimates (median LLGI=163.9) was slightly better than that calculated with *Oeffner* estimates (median LLGI=160.1) (Figure 4). However, a larger improvement was observed in the case of NMR models, where the median LLGI was 7.4 for calculations using the *Oeffner* estimates based on X-ray models and 14.7 using the new values derived for NMR models. The distribution of LLGI values for NMR models has also become much narrower using the new VRMS estimates (Figure 4). Note that few NMR models in our tests yield an LLGI score of 60 or more, which would normally indicate a correct solution, but the new LLGI values have been brought into a range that should help to enrich a pool of potential solutions with correct solutions (McCoy *et al.*, 2017). It should be noted that the calculations reported here used individual NMR models in order to calibrate the VRMS estimates, but in a real molecular replacement search one would use the whole ensembles, which would improve the results.

### 3.4. Comparative analysis of X-ray and NMR models

Our error estimates show why molecular replacement with NMR models is a challenge, as NMR models have much higher estimated errors than comparable X-ray models. To compare model quality over the whole range of sequence identity, for structures of the typical size addressed by NMR, we supplemented our data set with all available models between 60 and 100% sequence identity for targets in our data base between 125 and 175 residues in size, adding 444 X-ray models and 20 NMR models. For this size range, we found that using an NMR model with 90-100% sequence identity is equivalent to using an X-ray model of about 20-30% sequence identity (Figure 5). The data in this figure can be approximated

reasonably well by assuming that the NMR models differ in having an additional independent error component with a standard deviation of about 1.25 Å. This error component dominates across the sequence identity distribution.

## 4. Discussion

The *Oeffner* estimation of VRMS for X-ray models was systematically over-estimating the errors when the sequence identity was less than 30%. This artefact appears as a shoulder in the distribution of the ratio between refined and estimated VRMS (Figure 3 and Figure 5b from Oeffner *et al.*, 2013). Inspection of the cases populating this shoulder shows that this is due to limitations of using sequence identity to measure sequence similarity between distant homologs.

After the target and model sequences are optimally aligned, sequence identity represents a binary (True/False) score for each position in the alignment, which becomes a rather coarse measure for distant homologues with low sequence identity. Sequence identity also fails to distinguish between conservative and non-conservative substitutions. Hence we considered 20 matrix scores, listed in Table 1 and discussed in the review by Vogt *et al.,* (1995), which were expected to give a sensitive assessment of sequence similarity between homologs with less than 50% sequence identity. When we consider the full range of sequence identity (10 to 100%), BLOSUM30, BLOSUM35, BLOSUM40, BLOSUM45 (Henikoff & Henikoff, 1992), Benner22, Benner 74 (Benner *et al.*, 1994), and Gonnet scores (Gonnet *et al.*, 1992) are all strongly correlated to VRMS, with similar correlations of -0.70 to -0.71 Sequence identity gives a slightly weaker correlation of -0.67 (Table 2). However, looking at progressively lower levels of sequence identity, where MR is more challenging, some scoring matrices start to perform better. The Benner22, Benner74 and Gonnet scores all yield a correlation of -0.38 for models with sequence identity below 30%; for models with sequence identity below 20%, the Gonnet score gives a correlation of -0.15, slightly better than -0.14 for Benner74 and -0.11 for Benner 22. Our observations agree with an earlier finding that the Gonnet score is one of the top 3 matrices to assess sequence similarity among distant homologs (Vogt *et al.*, 1995). By replacing sequence identity with the Gonnet score, we have addressed the systematic over-estimation of errors in the distant homology regime.

While we were expecting to find a correlation to the resolution of the *model*, we were surprised to find *target* resolution instead to be correlated to the VRMS. Several other target properties such as asymmetric unit volume, Wilson B, and Matthews coefficient are also correlated to VRMS, but they are all correlated to each other and to target resolution. Once the resolution of the target was accounted for in the VRMS estimation, there were no residual correlations to these other target properties. This finding indicates that a higher RMSD should be expected if the crystal has diffracted to lower resolution. It could be explained by noting that crystals diffracting to lower resolution are intrinsically less well-ordered and possess a larger number of conformational states, which are explained poorly by a single

model. Similar conclusions have been drawn in the context of the gap between $R_{cryst}$ and $R_{merge}$ (Holton *et al.*, 2014).

Of the properties considered for evaluating model quality, resolution of the model, $R_{free}$, clashscore and *MolProbity* score were all correlated with VRMS, with *MolProbity* score giving the highest correlation. These measures were all correlated to each other, and once the influence of *MolProbity* score had been accounted for, there were no residual correlations with other properties of the model. Considering that *MolProbity* score (Chen *et al.*, 2010) combines contributions from clashscore, Ramachandran outliers and rotamer outliers, it is surprising that *MolProbity* is a significantly better predictor than clashscore, even though the correlations with Ramachandran and rotamer outliers are small. This presumably indicates that *MolProbity* score nonetheless integrates the influence of all three measures to assess the quality of model building and refinement better than any of the measures on its own.

The properties correlated to VRMS in the case of X-ray models were also found to be correlated to VRMS for NMR models. However, the relative importance of these factors differs. For the X-ray case, the most important factors were sequence similarity measured by Gonnet score, followed by number of residues in the model, resolution of target, and *MolProbity* score of the model. However, the number of residues in the model is the dominant factor for the NMR case with a correlation of 0.5, followed by Gonnet score, resolution of the target, and NMR ensemble consistency (measured as median RMSD between the models). Using the X-ray equation to estimate VRMS for NMR models will systematically underestimate the errors (Figure 3) leading to sub-optimal molecular replacement calculations, so a separate non-linear least squares fit was performed for NMR models.

With the new functional forms, we have achieved better accuracy and a better (more symmetrical and unimodal) distribution of errors for the estimates. The new estimates perform better for both X-ray and especially for NMR models.

Representing and querying highly interconnected data as a graph simplifies data analytics. The graph database has enabled us to overcome redundancies in the data and provided an easy way of extending the existing X-ray data along with the NMR data. It provided a platform to compare results from several trials of molecular replacement runs quickly and consistently. Further extension of the data in future, for example to include cryo electron microscopy related data, would also be possible.

By including properties of the target in the error estimates, we are pushing the boundaries of molecular replacement by personalising the model for a given data set. The data-driven model generation will pave the way for handling complex molecular replacement search strategies for structures with multiple domains or subunits.

The new VRMS estimates will be available as part of the *phaser.voyager* pipeline to run the new version of *Phaser*, *phasertng* (McCoy *et al., this issue*), which is currently under development.

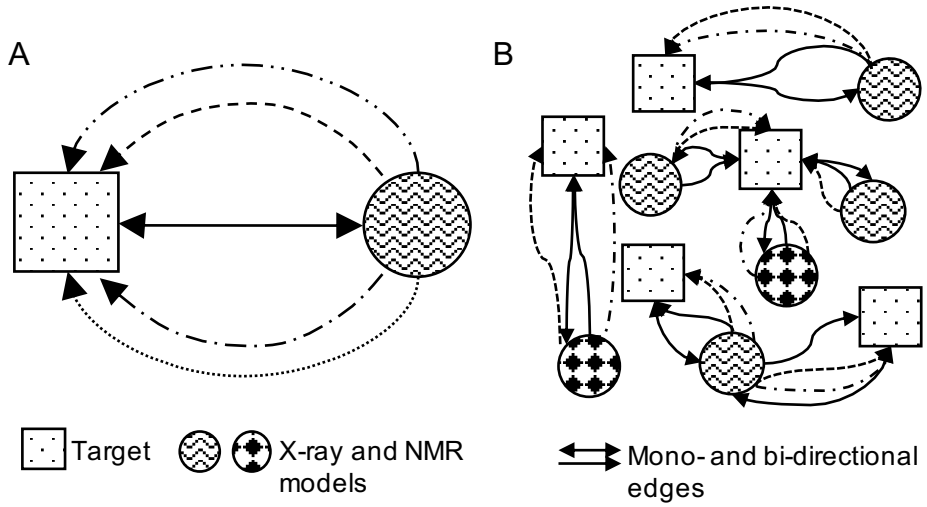


**Figure 1** Schematic representation of graph database. Targets and models are represented as square and circular nodes while an edge connecting two nodes represents a relationship between target and model node. A) Two types of edge can connect a target-model pair: 1) a uni-directional edge defines a single instance of a molecular replacement trial where a model was used to determine the target structure. The four different uni-directional edges represent four different trials of molecular replacement, for instance using data to different resolution limits. 2) a bi-directional edge defines properties associated with sequence similarity measures. More than one uni-directional edge exists between a target-model pair if more than one molecular replacement trial was carried out. B) presents an overview of a small graph database to show interconnections between the nodes. A single PDB entry could be used to determine two different targets; in which case the properties associated with processing the model, such as *MolProbity* score of the processed model are stored as part of the edge property. There are also examples where a single target could be determined with multiple independent models.

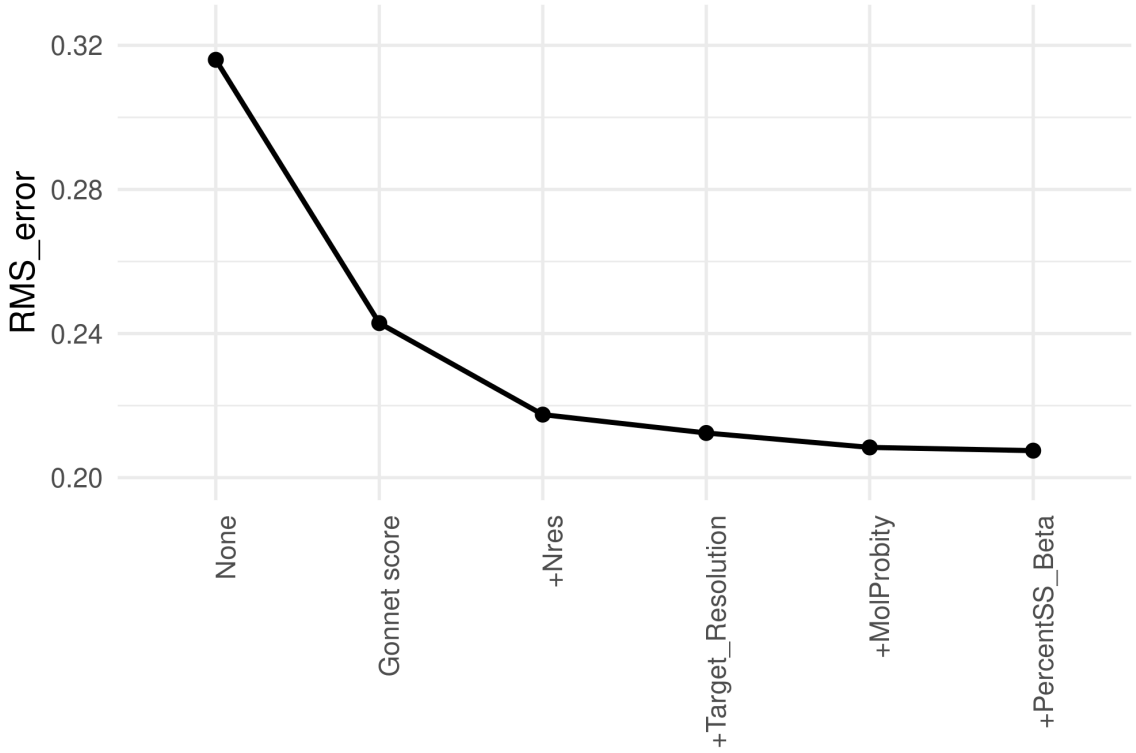


**Figure 2** RMS error in estimated VRMS as new properties are added to the prediction. Before any properties had been included ("None"), the RMS error was the RMS deviation of the refined VRMS values from their mean for all calculations.

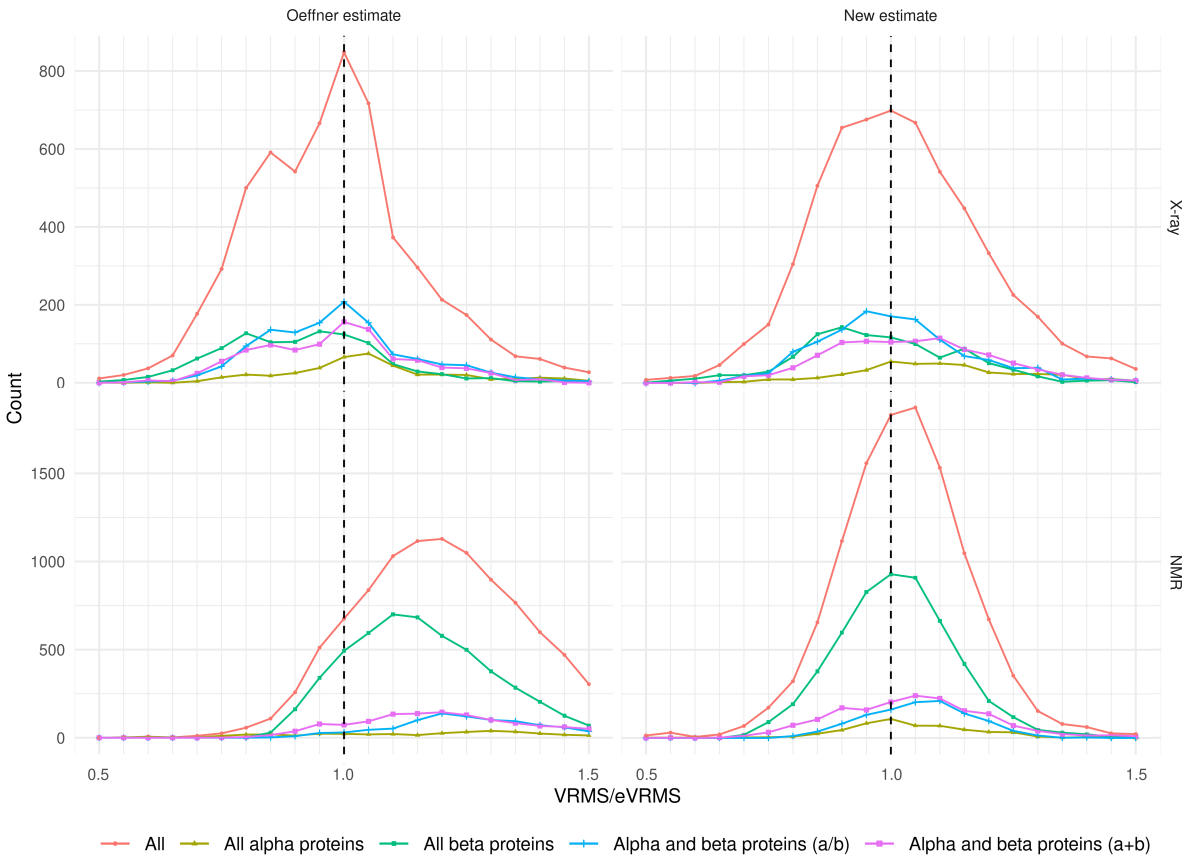


**Figure 3** Frequency distribution of refined over estimated VRMS ratios from the curated dataset as a function of SCOP class. A red line represents all cases. An ideal distribution should be Gaussian, with the lowest possible variance, and centred on 1 (represented by a black dashed line). **X-ray case**: The *Oeffner* estimate has a shoulder, which is not present in the new X-ray estimate. **NMR case:** The distribution for the *Oeffner* estimate based on X-ray data is shifted to the right, indicating that errors are systematically underestimated when applied to models derived by NMR. The new estimate based on NMR data has a symmetrical distribution centred around 1.

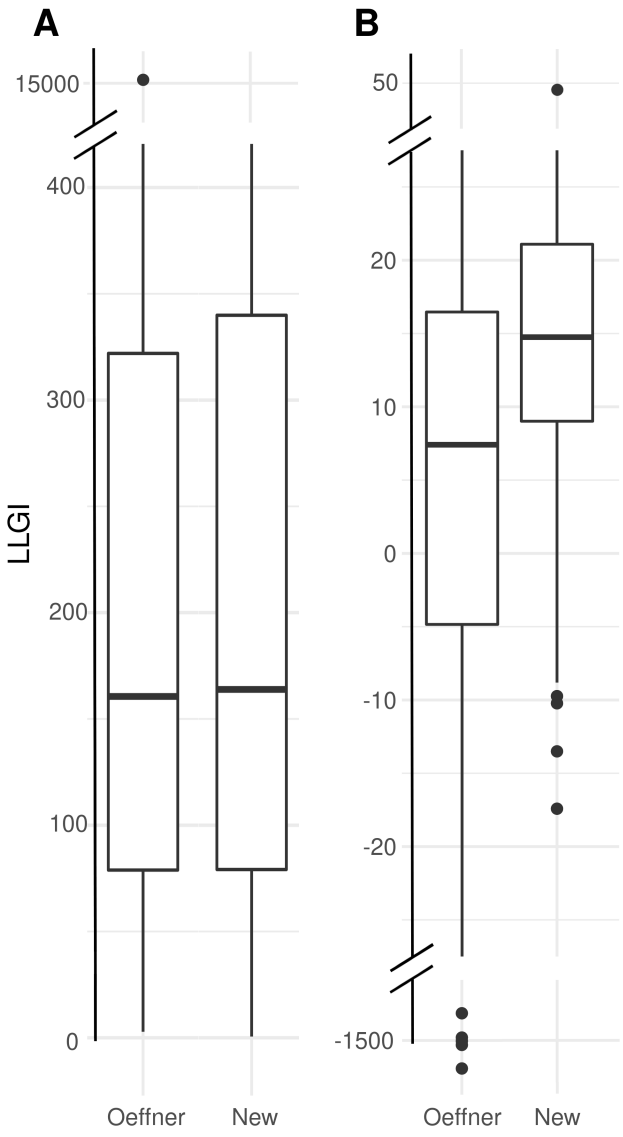


**Figure 4** Calculation of LLGI starting with *Oeffner* and new estimates of VRMS performed without VRMS refinement. A) Values for X-ray models. B) Values for NMR models. A limited range of LLGI values (along with the most extreme outliers) is displayed for the sake of clarity.

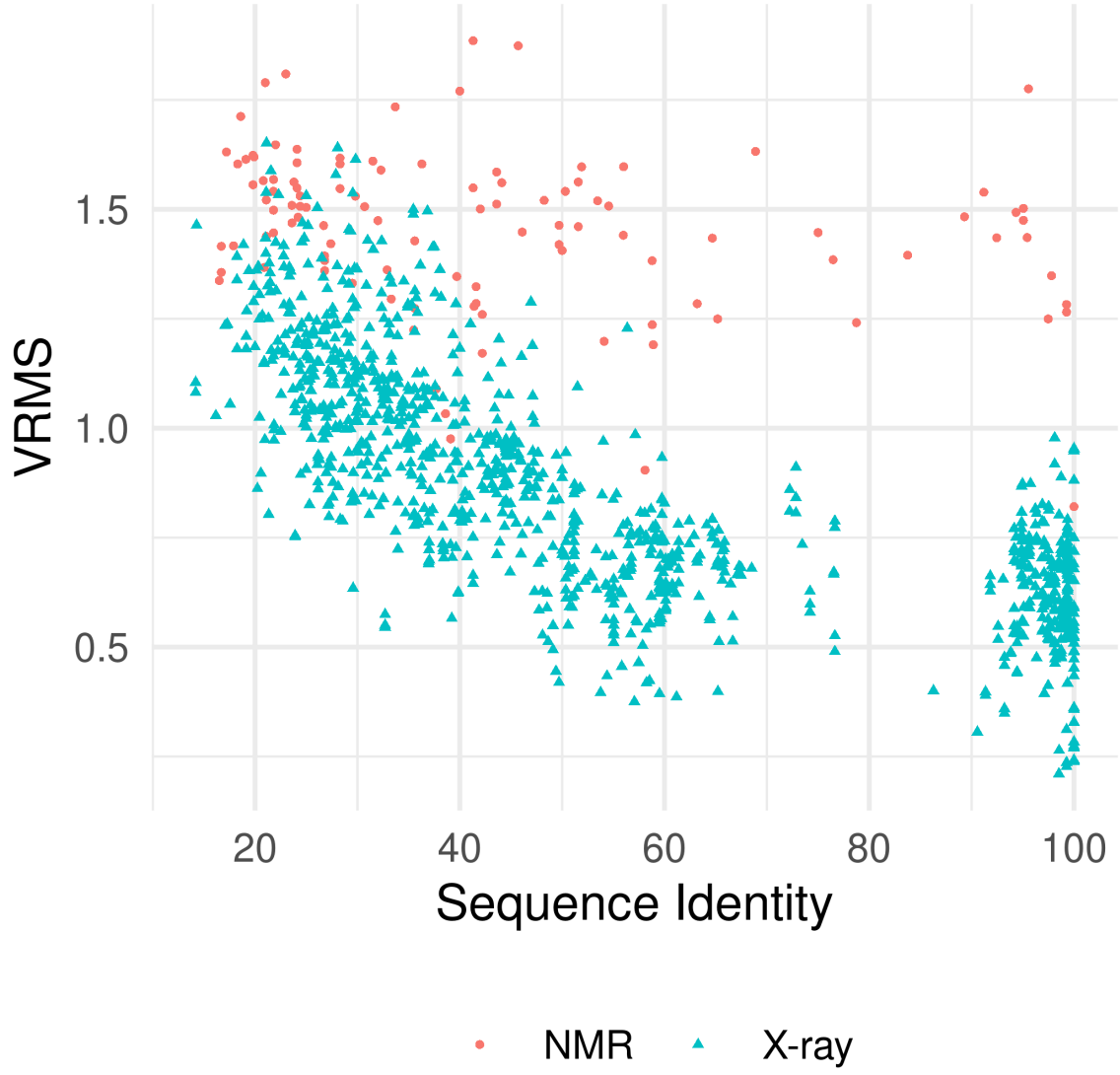


**Figure 5** Comparative analysis of errors between X-ray and NMR models of size 150±25 residues. Though the Gonnet score was used to estimate VRMS (y-axis), sequence identity (x-axis) is provided for ease of comparison.

**Table 1** List of properties considered in the study. The sequence similarity measures have been discussed in a previous review (Vogt *et al.*, 1995) and citations within.

| **Target Properties** | **Model properties** | **Sequence similarity measures** |
|---|---|---|
| Crystal Parameters: ASU volume, Unit cell dimension, Space group, Matthews coefficient, Crystal system, Polar space group<br><br>Data Parameters: Resolution, Wilson B, Merging statistics<br><br>Protein properties: Number of residues, SCOP Class | Validation Parameters: Ramachandran properties, Clashscore, Rotamer Outliers, *MolProbity* score, RMS Angles, RMS Bonds, Cβ-deviations, R-factors‡<br><br>Data properties: Resolution‡, completeness of resonance assignments†, Ensemble consistency†, Number of conformers deposited†, Number of conformers calculated†, Field strength†<br><br>Protein properties: Number of residues, Molecular weight, Non-sphericity, helix and sheet content<br><br>Deposition date | Sequence identity, PAM250, PAM300, BLOSUM30, BLOSUM35, BLOSUM40, BLOSUM45, BLOSUM65, Benner6, Benner22, Benner74, Feng, Genetic, Gonnet, Johnson, Levin, McLach, Miyata, Rao, Risler, Structure based |

‡ Properties specific to X-ray models. †Properties specific to NMR models. Ensemble consistency is measured as median RMSD between the models in an NMR ensemble.

**Table 2** Correlation of properties to X-ray VRMS term. Residual correlation is the correlation between the property and the difference between the estimated VRMS and the refined VRMS estimated either with the *Oeffner* equation (2) or the new equation (3)).

| Property | Correlation to VRMS | Residual correlation to VRMS | |
|---|---|---|---|
| | | ***Oeffner* estimate** | **New estimate** |
| **Number of residues of model** | 0.43 | 0.10 | 0.00 |
| **Sequence Identity** | -0.67 (-0.33*) | 0.00 | 0.00 |
| **Gonnet score** | -0.71 (-0.41*) | -0.16 | -0.03 |
| **Target Resolution** | 0.26 | 0.24 | 0.00 |
| ***MolProbity* score of model** | 0.16 | 0.18 | -0.02 |
| **Percent alpha-helix** | 0.20 | 0.19 | 0.10 |
| **Percent beta-sheet** | -0.14 | -0.16 | -0.13 |

*correlation for a subset of cases with <30% sequence identity

**Table 3** Correlation of properties with VRMS for the case of NMR models. Residual correlation is the correlation between the property and the difference between estimated and refined VRMS terms.

| Property | Correlation to VRMS | Residual correlation to VRMS | |
|---|---|---|---|
| | | ***Oeffner* X-ray estimate** | **New estimate** |
| **Number of residues of model** | 0.56 | 0.28 | 0.06 |
| **Gonnet score** | -0.38 | 0.40 | 0.00 |
| **Target Resolution** | 0.28 | -0.05 | -0.01 |
| **Median RMSD** | 0.22 | 0.14 | -0.02 |
| ***MolProbity* score of model** | 0.11 | 0.05 | 0.00 |
| **Percent alpha-helix** | 0.23 | 0.22 | 0.00 |
| **Percent beta-sheet** | -0.07 | -0.24 | -0.01 |

**Acknowledgements** This research was supported by funding from CCP4 (KSH), fellowship support from the European Union's Horizon 2020 research and innovation program under a Marie Skłodowska-Curie grant (MDS: 790122), a Wellcome Trust Principal Research Fellowship (RJR: grant 209407/Z/17/Z) and the NIH (grant P01GM063210 to RJR), which is gratefully acknowledged. We thank reviewers for their helpful comments.

**References**

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindah, E. (2015). *SoftwareX*. **1**–**2**, 19–25.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221.

Altschul, S. F. (1991). *J. Mol. Biol.* **219**, 555–565.

Baty, F., Ritz, C., Charles, S., Brutsche, M., Flandrois, J.-P. & Delignette-Muller, M.-L. (2015). *J. Stat. Softw.* **66**,.

Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1994). *Protein Eng. Des. Sel.* **7**, 1323–1332.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2013). *Acta Crystallogr. D. Biol. Crystallogr.* **69**, 2194–2201.

Bunkóczi, G. & Read, R. J. (2011*a*). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **7**, 303–312.

Bunkóczi, G. & Read, R. J. (2011*b*). *Comput. Crystallogr. Newsl.* **2**, 8–9.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. a, Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 12–21.

Chen, Y. W., Dodson, E. J. & Kleywegt, G. J. (2000). *Structure*. **8**, 213–220.

Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823–826.

Finn, R. D., Clements, J. & Eddy, S. R. (2011). *Nucleic Acids Res.* **39**, 29–37.

Fox, N. K., Brenner, S. E. & Chandonia, J.-M. (2014). *Nucleic Acids Res.* **42**, D304-9.

Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). *Science (80-. ).* **256**, 1443–1445.

Henikoff, S. & Henikoff, J. G. (1992). *Proc. Natl. Acad. Sci.* **22**, 10915–10919.

Holton, J. M., Classen, S., Frankel, K. A. & Tainer, J. A. (2014). *FEBS J.* **281**, 4046–4060.

Keegan, R. M. & Winn, M. D. (2008). *Acta Crystallogr. D. Biol. Crystallogr.* **64**, 119–124.

Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Crystallogr. D. Biol. Crystallogr.* **60**, 2240–2249.

Krissinel, E. (2012). *J Mol Biochem*. **1**, 76–85.

Mao, B., Guan, R. & Montelione, G. T. (2011). *Structure*. **19**, 757–766.

McCoy, A. J. *et al., Acta Crystallogr. Sect. D*.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Crystallogr.* **40**, 658–674.

McCoy, A. J., Oeffner, R. D., Wrobel, A. G., Ojala, J. R. M., Tryggvason, K., Lohkamp, B. & Read, R. J. (2017). *Proc. Natl. Acad. Sci. U. S. A.* **114**, 3637–3641.

Millán, C., Sammito, M. & Usón, I. (2015). *IUCrJ*. **2**, 95–105.

Montelione, G. T., Nilges, M., Bax, A., Güntert, P., Herrmann, T., Richardson, J. S., Schwieters, C. D., Vranken, W. F., Vuister, G. W., Wishart, D. S., Berman, H. M., Kleywegt, G. J. & Markley, J. L. (2013). *Structure*. **21**, 1563–1570.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Oeffner, R. D., Afonine, P. V, Millán, C., Sammito, M., Usón, I., Read, R. J. & McCoy, A. J. (2018). *Acta Crystallogr. Sect. D*. **74**, 245–255.

Oeffner, R. D., Bunkóczi, G., McCoy, A. J. & Read, R. J. (2013). *Acta Crystallogr. D. Biol. Crystallogr.* **69**, 2209–2215.

R Core Team (2018). *Software*.

Read, R. J. (1986). *Acta Crystallogr. Sect. A*.

Read, R. J. (2001). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **57**, 1373–1382.

Read, R. J. & McCoy, A. J. (2016). *Acta Crystallogr. Sect. D Struct. Biol.* **72**, 375–387.

Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 1288–1291.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D. & Higgins, D. G. (2011). *Mol. Syst. Biol.* **7**, 539.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). *Nucleic Acids Res.* **22**, 4673–4680.

Vogt, G., Etzold, T. & Argos, P. (1995). *J. Mol. Biol.* **249**, 816–831.

Wickham, H. (2016). ggplot2 - Elegant Graphics for Data Analysis Springer.

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N. & Alva, V. (2018). *J. Mol. Biol.* **430**, 2237–2243.